ChinaXiv合作期刊

知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

多任务环境下融合迁移学习的新冠疫情新闻要素识别研究

赵梓博[1,2] 王昊[1,2] 刘友华[1] 张卫[1,2] 孟镇[1,2]
1. 南京大学信息管理学院 南京 210023
2. 江苏省数据工程与知识服务重点实验室 南京 210023

摘要：[目的/意义]在新冠疫情背景下，提出多任务环境下融合迁移学习的疫情新闻要素识别方法，向公众提供面向应急事件的知识服务。[方法/过程]首先，通过多任务识别新闻要素：基于规则识别时间要素；并融合模型迁移与深度学习方法，构建跨领域的要素识别模型。在此基础上，构建疫情新闻要素的关联数据，以知识图谱的方式展示各要素之间的关联关系。[结果/结论]实验结果表明，除药物外的新闻要素的识别 F1 值均在 80% 以上，说明融合迁移学习的模型能够取得较优的识别效果；并且，关联数据知识图谱能够直观显示新闻的重点要素及新闻的主要内容。综上所述，提出的方法能够有效识别新冠疫情新闻要素，从而帮助新闻读者准确、高效地获取新闻中的重要信息。

关键词：多任务 迁移学习 新冠 新闻要素识别 命名实体识别 冷启动
分类号：TP391.1; TP181; G202
DOI：10.13266/j.issn.2095-5472.2021.001

引用格式：赵梓博，王昊，刘友华，等. 多任务环境下融合迁移学习的新冠疫情新闻要素识别研究 [J/OL]. 知识管理论坛，2021, 6(1): 2-13[引用日期]. http://www.kmf.ac.cn/p/235/.

1 引言

自 2020 年初，官方正式通报新型冠状肺炎（以下简称“新冠”）存在“人传人”现象以来，社会公众愈发关注新冠疫情的相关新闻动态。新冠疫情新闻对于帮助公众了解疫情动态、防疫方法等知识具有重要意义。然而，数量呈爆炸式增长的新闻报道给公众带来了一定程度的心理压力和阅读负担。因此，有必要快速、准确地提取新闻报道中的关键要素，帮助公众获取并理解新闻的主要内容，并为进一步构建疫情新闻知识图谱[1]提供数据支撑，为自动生

基金项目：本文系国家社会科学基金重大招标项目“情报学学科建设与情报工作未来发展路径研究”（项目编号：17ZDA291）和南京大学博士研究生创新研究项目“基于知识图谱的医学信息挖掘与推荐研究”（项目编号：CXYJ21-69）研究成果之一，并受江苏青年社科英才和南京大学仲英青年学者等人才培养计划的支持。
作者简介：赵梓博（ORCID:0000-0001-7487-5756），硕士研究生；王昊（ORCID:0000-0002-0131-0823），教授，博士生导师，通讯作者，E-mail: ywhaowang@nju.edu.cn；刘友华（ORCID:0000-0002-5859-9795），教授，硕士生导师；张卫（ORCID:0000-0003-4050-7255），博士研究生；孟镇（ORCID:0000-0002-1130-4996），硕士研究生。
收稿日期：2020-12-09 发表日期：2021-01-15 本文责任编辑：刘远颖

知识管理论坛，2021（1）:2-13
DOI: 10.13266/j.issn.2095-5472.2021.001

成疫情新闻关键词[2]、自发推送疫情新闻[3]等工作奠定基础。

新闻要素通常包括时间、人物、地点、机构4类基本要素，而新冠疫情新闻在此基础上还涉及疾病名称、发病症状、药物名称、诊断或治疗方法等医学要素，因此新冠疫情新闻要素识别需要对跨领域的多个类别的要素进行识别，这就涉及到多任务、多过程的要素识别。时间要素的表述形式具有较强的规律性，基于规则模板能够较准确地对其进行识别[4-5]，因此笔者采取基于规则的要素识别方法识别时间要素；而对于人名、地名、机构名3类基本要素以及疾病、症状、药物、方法4类医学要素，利用基于现有深度学习模型的命名实体识别（Named Entity Recognition, NER）方法进行识别，但是，疫情新闻作为一类新型应急信息资源，目前该领域尚存在缺乏供NER模型训练的标注数据这一数据冷启动问题，为此，笔者引入迁移学习思想，设计了跨领域迁移的实体识别模型。

笔者基于NER领域较为成熟的BERT-BiLSTM-CRF三层结构模型，分别利用MSRA数据集和医学领域数据集训练可迁移的NER模型，并将该模型应用于新冠疫情新闻领域的要素识别。最后，通过构建基于共现频次的要素关联数据，以知识图谱的方式可视化地展现疫情新闻要素间的关联关系，从而清晰、直观地揭示疫情新闻的主要内容。

2 近期相关研究

新闻文本要素的识别与提取是信息抽取领域的研究热点之一，在以往的实践中大多采用基于词典[6-7]、基于规则[8-9]或基于统计机器学习[10-12]的方法进行。近年来，随着深度学习研究的逐渐成熟，基于深度神经网络的命名实体识别也成为新闻要素识别的重要支撑技术[13-15]。相比传统机器学习算法，深度学习模型具有网络层数更深、学习特征更加复杂且无需人工构建特征等优势[16]。近年来提出的双向长短时记忆网络（Bidirectional Long Short-Term Memory, BiLSTM）[17]通过叠加句子在顺序和逆序方向的隐层表示，能够极大程度地揭示句中实体的依赖关系，因此被广泛应用于NER任务。研究表明，将BiLSTM与条件随机场（Conditional Random Field, CRF）相结合能够有效提高模型效果[18]。由谷歌AI团队于2018年发布的字表示模型BERT[19]，刷新了11项自然语言处理任务的记录。将BERT中文预训练模型（BERT-Base, Chinese）与识别效果较好的BiLSTM-CRF模型结合，被多项研究证实能够取得中文NER的最优效果[20-22]。

深度学习模型由于学习能力极强，易出现过拟合问题，因此需要庞大规模的标注数据作为训练集，而部分领域由于缺乏足够的训练数据而存在数据冷启动问题。为了解决这一问题，迁移学习（Transfer Learning）[23]的概念应运而生，其将在源领域学习到的知识应用于与源领域不同但相关的目标领域的任务中，利用源领域的标注数据训练可供目标领域应用的模型。迁移学习主要包括基于实例、基于特征和基于模型的迁移学习，基于实例的迁移学习的原理是将与目标领域实例相似的源领域样本加入训练集，以扩充数据量[24-25]；基于特征的迁移学习是指通过一定的方法，获取并利用源领域与目标领域之间共同的特征表示，从而实现表示层面的迁移[26-27]；基于模型的迁移学习是将基于源领域数据训练的模型及参数迁移至目标领域[28-29]。模型迁移学习基于大量源领域数据训练得到具有较强泛化能力的预训练模型，能够较好地适应目标领域的数据分布，从而取得较优的迁移效果，因此被广泛应用于NER领域。M. Al-Smadi等构建了基于迁移学习的多语言通用语句编码器，并将其应用于复杂阿拉伯语语境下的实体识别任务[30]；刘宇飞等将公共领域源知识迁移至科学领域，进而对专利文献中的科学术语进行识别[31]；孔祥鹏等提出基于迁移学习的联合深度模型，通过共享网络隐藏层以及BP算法微调参数的方法训练跨语言迁移模型，有效提升了维吾尔语NER任务的成绩[32]。

上述研究构建的迁移学习模型均取得了较好的实体识别效果，但是尚未考察以医学论文语料作为源领域训练数据的模型效果。考虑到新冠疫情新闻是一种面向当下应急事件的即时信息资源，领域内尚缺乏大规模的标注语料，笔者融合模型迁移与深度学习方法，以医学论文文本作为源领域数据集，基于学习效果较优的 BERT-BiLSTM-CRF 三层结构模型，训练实体识别模型，并将模型应用于疫情新闻要素的识别。

3 数据与方法

3.1 数据来源及预处理

笔者选取澎湃新闻发布的新冠疫情专题系列报道作为新冠疫情新闻文本的数据来源。由于澎湃新闻在我国新闻媒体网站排行榜排名居于前列 [33]，其文章质量较高，用词和句法较为规范和标准，因此适用于新闻要素抽取。基于模型迁移学习的思想，笔者确定以下两个源领域训练数据集：①微软亚洲研究院（MSRA）数据集，是中文 NER 任务的常用数据集，其语料含 27 000 余个句子，在本研究中将其用于识别人名、地名、机构名 3 类基本要素的基本要素识别模型的训练；②医学文本数据集，来源为中国知网平台新冠相关主题的中文医学论文题录数据，通过对论文题录数据进行处理后获得，其语料含 12 000 余个句子，用于识别疾病、症状、药物、方法 4 类医学要素的医学要素识别模型的训练。源领域数据集采用 IOB 格式进行实体标注，B 表示对应类别实体的起始字符，I 表示实体中的其他字符，O 表示非实体字符，如 B-PER 表示人名实体的起始字符，I-METHOD 表示方法实体中的非起始字符等。

笔者采用半监督的处理方法获得带标签的医学文本数据集，具体处理过程如下：①以“SU=’新冠’ + ‘新型冠状病毒’ + ‘武汉肺炎’ + ‘2019-ncov’ +’covid-19’”作为检索式，使用中国知网专业检索功能，搜索医药卫生科技分类下发表时间在“2020-02-01”后的中文论文，将检索结果显示的 6 000 条论文题录数据批量下载并保存；②提取题录数据中的关键词字段，人工对关键词进行实体类别标注，共得到 530 个标注后的关键词数据；③使用知网（Hownet）近义词词典，结合人工补充的方式，将原词的近义词标注为与原词相同的类别并补充入关键词集，扩充后的关键词集包含 607 个关键词；④提取题录数据中的全部摘要字段，通过最大匹配算法，使用标注关键词集匹配摘要文本中的句子，从而生成包含 12 000 余个含医学实体句子的医学文本语料。应用这种处理方法，只需要人工标注少量关键词，便能够匹配获得大量包含实体的句子，大大减少了人工标注的时间开销。

3.2 研究框架

为实现新冠疫情新闻要素的自动化识别及抽取，笔者设计了研究框架，见图 1。①首先，进行数据集的准备和预处理工作。分别收集 MSRA 数据集、医学论文题录数据以及新冠疫情新闻文本数据，然后人工标注医学论文题录数据中关键词的实体类别，并拓展关键词数量，随后利用拓展后的关键词集匹配论文摘要集中的句子，得到带有训练标签的医学文本数据集。②基于源领域数据集训练迁移要素识别模型。使用 BERT-BiLSTM-CRF 三层结构模型，分别基于 MSRA 数据集和医学文本数据集训练得到能够识别人物、地点、机构要素的基本要素识别模型 COV19News-Base 和能够识别疾病、症状、药物、方法要素的医学要素识别模型 COV19News-Med，并抽取原数据集中一定比例的样本作为测试集，以检验模型的识别效果。③将要素识别模型应用于新冠疫情新闻文本领域的要素识别。人工标注新冠疫情新闻文本中的部分句子作为目标领域测试集，分别检验将模型 COV19News-Base 和模型 COV19News-Med 应用于新冠疫情新闻要素识别的迁移效果。④最后，基于新闻要素构建要素关联图谱。使用 COV19News-Base 和 COV19News-Med 的模型组合抽取大量疫情新闻文本要素，结合基于

知识管理论坛，2021（1）:2-13
DOI: 10.13266/j.issn.2095-5472.2021.001

规则抽取的新闻时间要素，构建新冠疫情新闻要素关联数据，并以知识图谱的形式展现各要素之间的关联关系，以达到直观揭示新闻主要内容的目的。

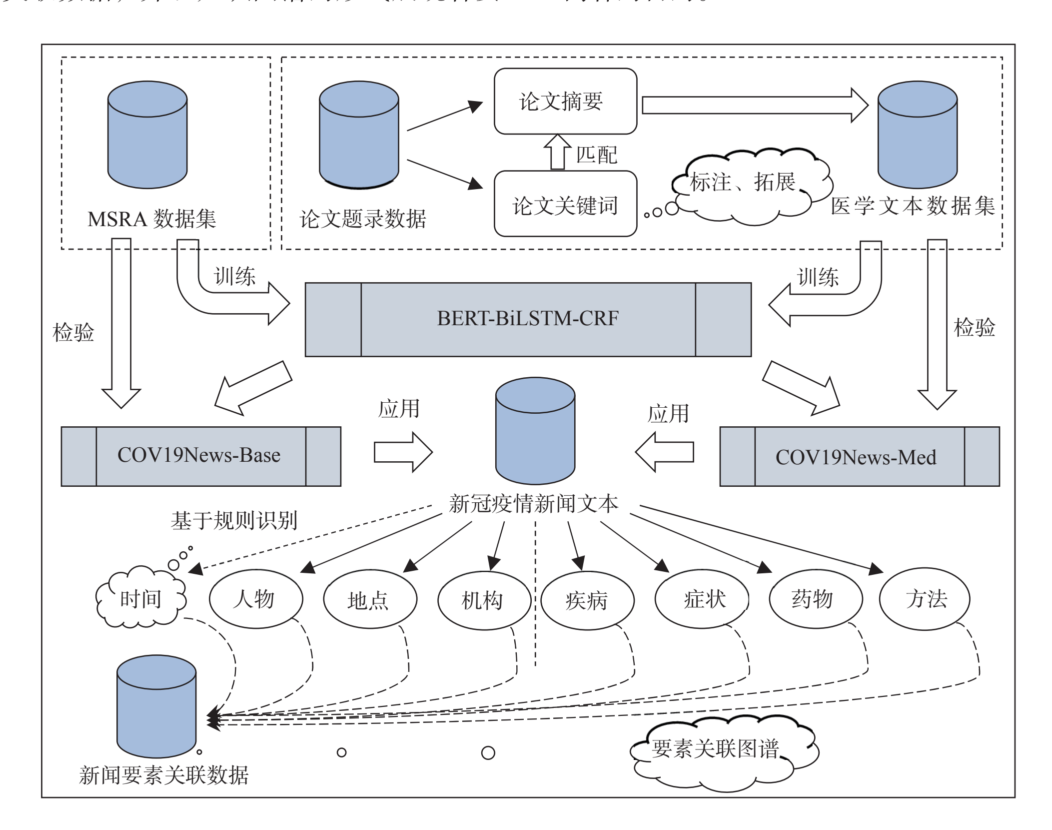


图 1 研究框架

基于此，笔者将主要解决以下 3 个重要问题：

（1）多类别要素的识别问题。将划分多个要素识别任务，基于命名实体识别和规则识别方法，分别对新冠疫情新闻中的基本要素、医学要素与时间要素进行识别。

（2）数据冷启动问题。引入模型迁移学习，利用源领域充足的标注数据训练可迁移的 NER 模型，并将其应用于疫情新闻领域的要素识别，从而解决了目标领域标注数据不充足的问题。

（3）疫情新闻要素的利用问题。将提出的要素识别方法应用于大量无标签的疫情新闻文本，并将识别的要素及要素间的共现关系以疫情新闻要素关联数据的形式存储。基于此，进一步以要素关联图谱的形式可视化展现要素间的关联关系，从而揭示疫情新闻的主要内容。

3.3 新冠疫情新闻要素分类

笔者试图实现 8 类疫情新闻要素的自动识别和抽取，8 类要素的名称及示例见表 1。其中，时间、人物、地点、机构 4 类要素是描述新闻内容的基本要素。此外，新冠疫情主题的新闻文本往往还包含疾病名称、发病症状、药物名称、诊断或治疗方法的名称等医学要素。对于具体识别哪些类别的医学要素，可借鉴前人研究的经验。在 2019 年全国知识图谱与语义计算大会（CCKS）医疗命名实体识别任务中，医疗命名实体被划分为 6 类：疾病和诊断、检查、检验、手术、药物、解剖部位 [20]；2017 年 CCKS 定义了 4 类医学实体：身体部位、症状和体征、检查和检验、疾病和诊断 [34]；赵青等、夏光辉等将医疗实体划分为疾病、症状、检查、治疗 4 类 [35-36]。由上述研究总结，医学实体总共包括 5 类：疾病名称、症状体征、

药物、检查和治疗方法以及身体部位。但身体部位实体在新闻领域语境下往往具有除患病部位以外的含义，如“握手言和”中的“手”“嘴上说说”中的“嘴”等并非指代患病部位，不属于描述新闻内容的关键要素，因此识别身体部位实体对提取新闻要点的意义不大。综上所述，笔者最终确定将疾病、症状、药物、方法 4 类要素作为待识别的医学要素。

表 1 新冠疫情新闻要素类别及示例

要素序号	要素名称	例句	句中要素
1	时间	1月24日，湖北省启动重大突发公共卫生事件Ⅰ级响应	1月24日
2	人物	灿扬的咽痛没有缓解的迹象，杨雪涛也出现了咳嗽的症状	灿扬，杨雪涛
3	地点	英国也有医院出现了病人被感染的情况	英国
4	机构	大学毕业后，李文亮去到厦门眼科中心工作	厦门眼科中心
5	疾病	3月4日，新冠肺炎的海外传播发酵近两月	新冠肺炎
6	症状	5例病例均发烧，无畏寒，体温在36.5° C-38.0° C之间波动	发烧，畏寒
7	药物	药物方面，他们现在为密接者提供了连花清瘟胶囊	连花清瘟胶囊
8	方法	核酸检测会出现一定的假阴性率	核酸检测

笔者通过多个任务识别各类疫情新闻要素。对于除时间要素以外的 7 类要素，采取命名实体识别方法对其进行识别，基于 BERT-BiLSTM-CRF 模型分别训练基本要素识别模型和医学要素识别模型；对于时间要素，采取基于规则的识别方法，通过构建正则表达式，匹配并获取新闻文本中的时间要素。匹配时间要素的正则表达式模板如公式（1）所示：

pattern=[r"\d{4} 年 \d{1,2} 月 \d{1,2}[日 | 号]",
r"\d{1,2} 月 \d{1,2}[日 | 号]\"",
r"\d{1,2}[日 | 号]\"" 公式（1）

3.4 基于迁移学习的 COV19News 模型训练

由于疫情新闻领域尚缺乏可供 NER 模型训练的标注数据，笔者采用融合迁移学习的模型训练方法，分别基于 MSRA 数据集和医学文本数据集训练模型 COV19News-Base 和模型 COV19News-Med，并将上述模型应用于疫情新闻文本中各类要素的识别。为了检验不同模型的识别效果，分别对 MSRA 数据集和医学文本数据集进行训练集、测试集的划分，以供模型 COV19News-Base 和模型 COV19News-Med 在源领域的训练和检验；并从新闻文本中分别抽取并标注 100 个包含基本要素和医学要素的句子，作为模型的目标域测试集。

在进行模型训练前，对源领域训练集、源领域测试集和目标领域测试集中的实体数量进行统计，统计结果见表 2，其中模型 COV19News-Base 的源领域数据集为 MSRA 数据集，模型 COV19News-Med 的源领域数据集为医学文本数据集，两模型的目标领域测试集均为新闻文本中抽取的句子。从表 2 中可以发现，源领域数据集存在不同程度的实体分布不均衡现象，MSRA 数据集中地名实体明显多于人名和机构名实体，而医学文本数据集中疾病实体更远多于其他 3 类实体，这是由于来自医学论文的标注关键词集中大部分关键词属于疾病实体，主要包括新冠的大量别称，因此造成了匹配实体数量分布不均匀的问题。从目标领域测试集实体分布的角度看，人名、地名、机构名 3 类实体分布较为均匀，而医学实体中疾病实体仍然是出现频率最高的实体，这与新冠疫情新闻的特点有关（报道中包含较多新冠的指代与别称）。实体分布的不均衡是否会影响模型效果有待实验考证。此外，医学文本数据集的规模相对 MSRA 数据集较小，因此可供训练的实体数量相对较少，可能会对模型效果造成影响，具体有待后续探究。

ChinaXiv合作期刊
知识管理论坛，2021（1）:2-13
DOI: 10.13266/j.issn.2095-5472.2021.001

表 2 数据集中实体的数量统计情况

模型	实体类别	实体数量统计/个		
		源领域训练集	源领域测试集	目标领域测试集
COV19News-Base	人物	8 144	2 748	73
	地点	16 571	5 609	80
	机构	9 277	3 169	83
COV19News-Med	疾病	13 443	2 178	82
	症状	1 734	309	57
	药物	1 183	202	30
	方法	1 986	347	35

基于 BERT-BiLSTM-CRF 模型，使用上述训练数据分别训练模型 COV19News-Base 和模型 COV19News-Med。BERT 采用多层的双向 Transformer[37] 编码器结构，能够捕捉长距离上下文的语义特征，从而得到较为精确的文本向量；BiLSTM 采用二重逆序的 LSTM 网络，能够充分学习向量间双向的语义关系；CRF 则能够依照序列标签的约束规则，输出全局最优的标记序列。因此，采用 BERT-BiLSTM-CRF 模型进行模型训练，在模型表示层、网络层和输出层均能取得较优的学习效果，适用于 COV19News 模型的训练。模型训练完毕后，分别基于源领域和目标领域测试集对模型效果进行检验，检验结果见实验结果与分析部分。

3.5 疫情新闻要素的知识图谱构建

在利用上述模型实现对疫情新闻要素的识别和提取后，进一步构建疫情新闻要素的知识图谱，可视化展现要素间的关联关系。

考虑到疫情新闻要素之间存在关联关系，并且要素间的关联能够揭示新闻的主体事件，因此对新闻要素关联关系的挖掘有助于推断疫情新闻的主要内容，对读者理解新闻内容具有重要的意义。首先将整篇新闻文本划分为句子的集合，然后将在同一句子中出现的要素记为共现一次，由此计算两两要素的共现频次，以“要素 A- 要素 B- 共现频次”的格式保存为数据文件，作为疫情新闻要素的关联数据。疫情新闻要素关联数据描述了要素间的关联关系以及关联关系的强度，为疫情新闻要素知识图谱的构建提供了数据支撑。

疫情新闻要素知识图谱能够清晰、直观地展现要素关联及其强度，有助于读者定位新闻中的关键要素，进而推断新闻的主要内容。因此，基于新闻要素关联数据，以要素作为节点，两要素的共现频次作为两节点连线的权重，进一步构建疫情新闻要素的关联数据知识图谱。笔者使用网络分析软件 Gephi 绘制疫情新闻要素关联知识图谱，见图 2。由图 2 可知，新闻中与其他要素关联较为紧密的关键要素得到了突出显示，并且根据要素间的关联关系，读者能够联系各个要素，对新闻的主要内容进行推断。



图 2 使用 Gephi 绘制疫情新闻要素知识图谱演示

4 结果与分析

4.1 实验环境及模型参数设置

模型的训练、测试和迁移全部在装载 6GB 显存的 NVDIA GeForce RTX 2060 显卡、内存 16GB、操作系统为 Windows10 的个人计算机中进行，模型运行环境为 Python3.5 + Tensorflow1.12GPU 版，CUDA 版本为 10.2。BERT-BiLSTM-CRF 模型的部分参数如表 3 所示：

表 3 模型参数设置

序号	模型参数	参数含义	参数取值
1	Batch Size	一次送入训练模型的样本（字符）数	128
2	Segment Embedding Size	表示BERT句子嵌入的向量维度	20
3	Char Embedding Size	表示BERT字符嵌入的向量维度	100
4	Dropout Rate	BiLSTM网络的遗忘率	0.5
5	Learning Rate	BiLSTM网络的学习率	0.001
6	Optimizer	网络前向传播使用的优化器类型	adam
7	Max Epoch	迭代轮数	100

4.2 模型 COV19News-Base 的测试与迁移

笔者采用精确率（Precision, P）、召回率（Recall, R）以及二者的调和平均值（F1-measure, F1）评估模型的识别效果。对于通常包含多个单字的实体，当且仅当模型输出的实体标签序列与原标注序列完全相同时，记为正确识别实体，否则记为错误识别。在后续实验中，OP、OR、OF1 分别表示模型在源领域的 P、R、F1 值，TP、TR、TF1 分别表示模型在目标领域的 P、R、F1 值。

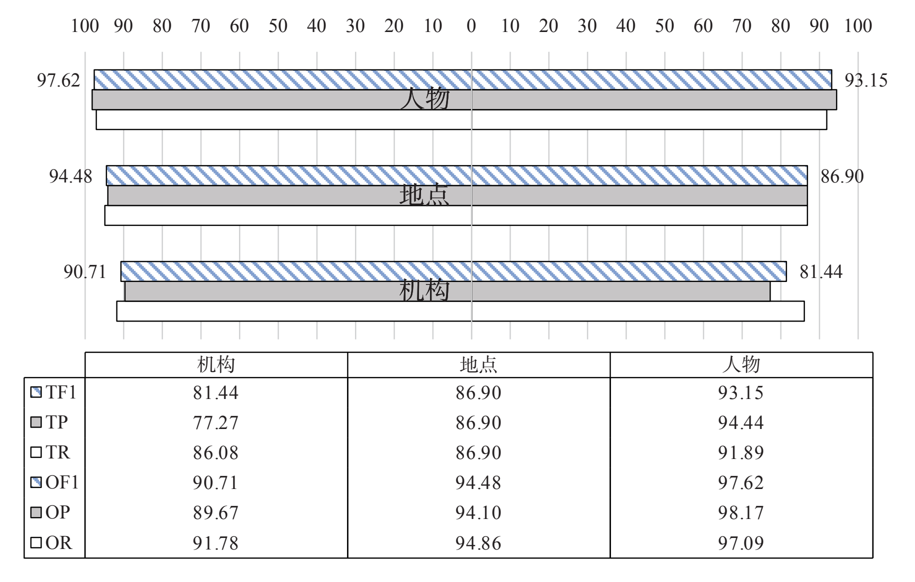
基于 MSRA 数据集训练模型 COV19News-Base，源领域和目标领域的测试集表现如图 3 所示。由图可知：①由于同领域的训练集和测试集的实体分布特征较为一致，因此模型在源领域测试集上表现出较优的识别效果，3 类实体的 F1 值均在 90% 以上。②模型迁移至目标领域后，3 类实体的识别效果均出现了不同程度的下降，但 F1 值仍能保持在 80% 以上。考虑到疫情新闻领域文本与 MSRA 数据集在实体分布上存在差异，迁移后模型识别效果的略微下滑符合预期。③对 3 类实体的识别效果进行相比，人物实体的识别效果最优，其次是地点实体，机构实体的识别效果最差。地点和机构实体的平均长度通常大于人物实体，其识别难度也相对更大，因此模型对不同实体的识别效果存在差异。④虽然地点实体在源数据集中的出现频率高于其他两类实体，但其识别效果并未更优，这说明训练集中实体的不均衡分布并未影响模型效果。

4.3 模型 COV19News-Med 的测试与迁移

复原模型的基础参数，基于医学文本数据集训练模型 COV19News-Base，源领域和目标领域的测试集表现如图 4 所示。可以发现：①模型在源领域测试集的表现仍然较优，4 类医学实体的识别 F1 值均在 90% 以上，表明 BERT-BiLSTM-CRF 框架具有较强的表征和学习能力，对于不同领域的数据均能够保持较好的拟合效果。②虽然医学文本数据集相较 MSRA 数据集规模较小，但在源领域测试集的表现并未落后，说明在数据规模量级达标的前提下，投入相对少量的样本也能使模型取得较好的训练结果，不会影响模型效果。③将模型迁移至目标领域后，各类实体的识别效果出现了不同程度的下滑，但除药物实体外，其他 3 类实体的 F1 值仍能保持在 80% 以上，较符合预期。识别效果下降是因为各类实体在目标领域测试集的召回率

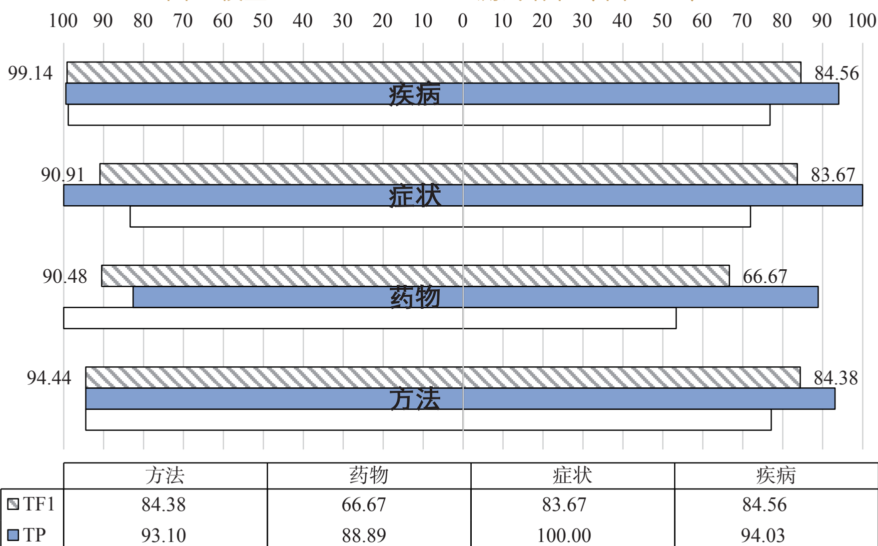
知识管理论坛，2021（1）:2-13
DOI: 10.13266/j.issn.2095-5472.2021.001

表现较差，可能因为医学论文文本与疫情新闻文本中医学实体的分布特征存在较大差异，导致模型迁移后的泛化效果不够理想，使得一部分目标领域中存在但未能被模型学习的实体难以被识别。尽管如此，迁移后的模型依然能保证较高的识别精确率。④在源领域数据集中，疾病实体的数量远超出其他 3 类实体，疾病实体在源领域和目标领域测试集的表现也最优，但在目标领域测试集的 F1 值与症状、方法两类实体相比差距已不明显。这表明，虽然极不均衡的实体分布可能会对某类实体在源领域的识别起积极作用，但是未必对该类实体在目标领域的表现产生较大影响，后者仍然与目标领域的实体分布特征有关。



	机构	地点	人物
TF1	81.44	86.90	93.15
TP	77.27	86.90	94.44
TR	86.08	86.90	91.89
OF1	90.71	94.48	97.62
OP	89.67	94.10	98.17
OR	91.78	94.86	97.09

图 3 模型 COV19News-Base 测试结果（单位：%）



	方法	药物	症状	疾病
TF1	84.38	66.67	83.67	84.56
TP	93.10	88.89	100.00	94.03
TR	77.14	53.33	71.93	76.83
OF1	94.44	90.48	90.91	99.14
OP	94.44	82.61	100.00	99.42
OR	94.44	100.00	83.33	98.85

图 4 模型 COV19News-Med 测试结果（单位：%）



上述实验结果表明，基于迁移学习方法训练得到的 NER 模型，对于目标领域疫情新闻要素的识别具有较好的效果。为展示所提出方法的识别效果，笔者在疫情新闻文本中随机选取多个包含多类要素的句子，使用模型 COV19News-Base 和模型 COV19News-Med 对其中要素进行识别，并基于时间要素的表述规则构建正则表达式模板，匹配并识别句子中的时间要素，最后将多个任务的识别结果汇总，部分结果如表 4 所示：

表 4 疫情新闻要素的识别结果举例

序号	示例	识别结果
1	那是2020年1月21日，新型冠状病毒感染的肺炎疫情正从武汉向全国蔓延，翁秋秋所在的湖北黄冈蕲春县距离武汉不过百余里，黄冈是武汉之外疫情最严重的地区	{2020年1月21日，新型冠状病毒感染，肺炎，武汉，翁秋秋，湖北，黄冈，蕲春县}
2	从1月26日出现乏力、咳嗽等症状以来，吴娟父亲一直在等待一个明确的诊断——如果没有确诊新型冠状病毒感染的肺炎（以下简称“新冠肺炎”），他就无法收治入院	{1月26日，乏力，咳嗽，吴娟，新型冠状病毒感染，肺炎，新冠肺炎}
3	华中科技大学公共卫生学院副院长徐顺清2月9日晚间在湖北省的新闻发布会上表示，新型冠状病毒侵害的部位主要是肺部，所以用核酸检测存在一定的假阴性，也就是有一部分病人没有检测出来，就是漏诊，这样可能造成一些传染源没有真正地被识别出来，有扩大的风险	{华中科技大学公共卫生学院，徐顺清，2月9日，湖北省，新型冠状病毒，核酸检测}
4	当奥地利方面3月4日通知该邮轮有奥地利男子感染新冠病毒时，“歌剧”号邮轮正停靠在雅典比雷埃夫斯港，出于安全考虑，“歌剧”号邮轮要求当时在比雷埃夫斯港上岸的所有乘客尽快返回船上	{奥地利，3月4日，新冠病毒，“歌剧”号邮轮，雅典，比雷埃夫斯港}
5	李文亮，男，35岁，武汉市中心医院眼科医生李文亮于2004年参加高考，从武汉大学临床医学七年制专业毕业后，先在厦门工作了三年，2014年回到武汉，一直在武汉市中心医院工作	{李文亮，武汉市中心医院，武汉大学，厦门，武汉}

4.4 新冠疫情新闻要素知识图谱的构建

基于上述疫情新闻要素的识别方法，提取新闻要素并构建要素关联数据，进而构建新冠疫情新闻要素的关联知识图谱。以一篇标题为《家属口述 | 一个“重症肺炎”患者的最后 12 天》的新闻报道为例，构建其要素知识图谱，如图 5 所示：



图 5 疫情新闻要素关联关系展示举例

由图 5 可知，该篇新闻主要涉及时间、人物、地点、机构、疾病要素，其中“翁秋秋”“武汉”“肺炎”为重要要素。结合要素关联情况推断，该篇新闻的主要内容为黄冈市民翁秋秋身患新冠，并于黄冈市中医院接受治疗。可见，疫情新闻要素的关联知识图谱能够有效帮助读者确定新闻重点以及推断新闻主要内容，因此有潜力成为面向新冠疫情突发事件的新型知识服务。

5 结论

笔者提出了一种多任务环境下融合迁移学习与深度学习技术的疫情新闻要素识别方法，为应急事件下公民的信息获取提供了可行的服务方案。首先，结合命名实体识别与规则识别方法，通过多个任务对多类别的新闻要素进行识别。同时，为解决疫情新闻领域数据冷启动的问题，采用模型迁移的解决方案，从而得到

ChinaXiv合作期刊
知识管理论坛，2021（1）:2-13
DOI: 10.13266/j.issn.2095-5472.2021.001

识别效果较好的跨领域要素识别模型。最后，将识别方案应用于大量新冠疫情新闻文本，基于识别到的新闻要素构建要素关联数据知识图谱，从而帮助新闻读者直观、快速地发掘新闻关键要素及主要内容。

通过对模型测试和迁移的效果进行比较，得到以下结论：① BERT-BiLSTM-CRF 三层结构模型适用于不同领域的命名实体识别任务，且源领域各类实体识别的 F1 值均在 90% 以上；②将模型由源领域迁移至目标领域后，模型的识别效果有下降趋势，但尚保持在可接受的范围内，大部分实体识别的 F1 值均在 80% 以上；③若源领域训练数据中实体分布极不均衡，可能导致对某类实体的过度学习，在源领域中对该类实体的识别效果远优于其他实体，但是否会影响目标领域实体的识别仍有待后续研究。

综上所述，笔者提出的基于迁移学习的要素识别方法对于新冠疫情新闻要素具有较优的识别效果。但本研究尚存在部分类别实体识别率较低等问题。在后续研究中，将重点考虑将实例迁移与模型迁移相结合，使训练域与目标域的实体分布更加接近，从而提升模型在目标领域的识别效果。

参考文献：

[1] 王岩 , 蒿兴华 , 薛鹏 . 基于共词分析和社会网络分析的关联数据知识图谱构建分析 [J]. 数字通信世界 , 2020(6):148-150.

[2] 陶洁 . 基于新闻文本的关键词提取 [D]. 武汉 : 华中师范大学 , 2019.

[3] 陶天一 , 王清钦 , 付聿炜 , 等 . 基于知识图谱的金融新闻个性化推荐算法 [J/OL]. 计算机工程 , 2020: 1-10 [2020-09-12]. https://doi.org/10.19678/j.issn.1000-3428.0057446.

[4] 裴韬 , 郭思慧 , 袁烨城 , 等 . 面向公共安全事件的网络文本大数据结构化研究 [J]. 地球信息科学学报 , 2019, 21(1):2-13.

[5] 吉雷静 . 面向网页文本的地理信息变化语义检测方法研究 [D]. 南京 : 南京师范大学 , 2013.

[6] 伏恺 . Web 新闻文本信息抽取与可视化研究 [D]. 济南 : 山东财经大学 , 2017.

[7] KRSTEV C, OBRADOVIC I, UTVIC M, et al. A system for named entity recognition based on local grammars[J]. Journal of logic and computation, 2014, 24(2):473-489.

[8] 杨建林 , 王文龙 . 公共卫生类突发事件的抽取研究 [J]. 情报理论与实践 , 2016, 39(4) :51-59.

[9] KUCUK D, YAZICI A. A hybrid named entity recognizer for Turkish[J]. Expert systems with applications, 2012, 39(3):2733-2742.

[10] SEKER G A, ERYIGIT G. Extending a CRF-based named entity recognition model for Turkish well formed text and user generated content[J]. Semantic Web, 2017, 8(5):625-642.

[11] 吴伟成 . 基于恐怖袭击事件语料库的时间短语抽取研究 [D]. 南京 : 南京大学 , 2016.

[12] CHASIN R, WOODWARD D, WITMER J, et al. Extracting and displaying temporal and geospatial entities from articles on historical events[J]. Computer journal, 2014,57(3):403-426.

[13] 李玉超 . 新闻事件地名实体识别和地图链接技术研究 [D]. 成都 : 电子科技大学 , 2020.

[14] WICHMANN P, BRINTRUP A, BAKER S, et al. Extracting supply chain maps from news articles using deep neural networks[J]. International journal of production research, 2020, 58(17):5320-5336.

[15] XU J G, GUO L X, JIANG J, et al. A deep learning methodology for automatic extraction and discovery of technical intelligence[J]. Technological forecasting and social change, 2019, 146 :339-351.

[16] 王昊 , 邓三鸿 , 朱立平 , 等 . 大数据环境下政务数据的情报价值及其利用研究——以海关报关商品归类风险规避为例 [J]. 科技情报研究 , 2020, 2(4):74-89.

[17] DONG X S, CHOWDHURY S, QIAN L J, et al. Deep learning for named entity recognition on Chinese electronic medical records: combining deep transfer learning with multitask bi-directional LSTM RNN[J]. PLOS one, 2019, 14(5):1-15.

[18] 肖连杰 , 孟涛 , 王伟 , 等 . 基于深度学习的情报分析方法识别研究——以安全情报领域为例 [J]. 数据分析与知识发现 , 2019, 3(10):20-28.

[19] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[EB/OL].[2020-09-12]. https://arxiv.org/abs/1810.04805.

[20] 李灵芳 , 杨佳琦 , 李宝山 , 等 . 基于 BERT 的中文电子病历命名实体识别 [J]. 内蒙古科技大学学报 , 2020, 39(1):71-77.

[21] 吴俊, 程垚, 郝瀚, 等. 基于 BERT 嵌入 BiLSTM-CRF 模型的中文专业术语抽取研究 [J]. 情报学报, 2020, 39(4):409-418.

[22] 刘忠宝, 党建飞, 张志剑.《史记》历史事件自动抽取与事理图谱构建研究 [J]. 图书情报工作, 2020, 64(11):116-124.

[23] YOSINSKI J, CLUNE J, BENGIO Y, et al. How transferable are features in deep neural networks? [EB/OL]. [2020-09-12]. https://arxiv.org/abs/1411.1792.

[24] 陈美杉, 夏晨曦. 肝癌患者在线提问的命名实体识别研究：一种基于迁移学习的方法 [J]. 数据分析与知识发现, 2019, 3(12):61-69.

[25] 李号号. 基于实例的迁移学习技术研究及应用 [D]. 武汉 : 武汉大学, 2018.

[26] 陈文珺, 杨佳佳. 基于共享知识迁移学习的跨领域推荐研究 [J]. 情报科学, 2020, 38(6):126-132.

[27] GLIGIC L, KORMILITZIN A, GOLDBERG P, et al. Named entity recognition in electronic health records using transfer learning bootstrapped neural networks[J]. Neural networks, 2020, 121 :132-139.

[28] KUNG H K, HSIEH C M, HO C Y, et al. Data-augmented hybrid named entity recognition for disaster management by transfer learning[J]. Applied sciences-basel, 2020, 10(12):1-17.

[29] 邵明锐, 马登豪, 陈跃国, 等. 基于社区问答数据迁移学习的 FAQ 问答模型研究 [J]. 华东师范大学学报 (自然科学版), 2019(5):74-84.

[30] Al-SMADI M, Al-ZBOON S, JARARWEH Y, et al. Transfer learning for Arabic named entity recognition with deep neural networks[J]. IEEE access, 2020,8:37736-37745.

[31] 刘宇飞, 尹力, 张凯, 等. 基于深度迁移学习的技术术语识别——以数控系统领域为例 [J]. 情报杂志, 2019, 38(10):168-175.

[32] 孔祥鹏, 吾守尔·斯拉木, 杨启萌, 等. 基于迁移学习的维吾尔语命名实体识别 [J]. 东北师大学报 (自然科学版), 2020, 52(2):58-65.

[33] 站长之家. 新闻媒体网站排行榜 [EB/OL]. [2020-09-30]. https://top.chinaz.com/hangye/index_news.html.

[34] 李飞, 朱艳辉, 王天吉, 等. 基于医疗类别的电子病历命名实体识别研究 [J]. 湖南工业大学学报, 2018, 32(4):61-66.

[35] 赵青, 王丹, 徐书世, 等. 一种基于 RNN 的弱监督中文医疗实体识别方法 [J/OL]. 哈尔滨工程大学学报 , 2020:1-10[2020-09-12]. http://kns.cnki.net/kcms/detail/23.1390.U.20200330.1522.002.html.

[36] 夏光辉, 李军莲, 邢宝坤, 等. 基于中文病例报告文献的医学诊疗命名实体识别研究 [J]. 医学信息学杂志, 2019, 40(6):54-59.

[37] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. [2020-09-12]. https://arxiv.org/abs/1706.03762.

作者贡献说明:

赵梓博：负责完成实验，撰写论文初稿；

王昊：指导研究思路，核查论文内容并提出修改意见；

刘友华：负责整理实验结果，审查异常数据指标并提出改进策略；

张卫：提供有关可视化方法、工具的指导建议，并参与修改终稿；

孟镇：负责修改终稿。

知识管理论坛，2021（1）:2-13
DOI: 10.13266/j.issn.2095-5472.2021.001

Research on Identification of COVID-19 News Elements based on Transfer Learning in Multi-task Environment

Zhao Zibo[1,2] Wang Hao[1,2] Liu Youhua[1] Zhang Wei[1,2] Meng Zhen[1,2]
[1] School of Information Management, Nanjing University, Nanjing 210023
[2] Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210023

Abstract: [Purpose/significance] Under the background of novel coronavirus pneumonia, this paper proposes a method of identifying COVID-19 news elements in multi-task environment based on transfer learning to provide knowledge services of emergency for the public. **[Method/process]** Firstly, multiple tasks were used to identify news elements: Time elements were identified based on rules; besides, a cross domain element recognition model was constructed by integrating model transfer and deep learning methods. On this basis, the associated data of COVID-19 news elements was constructed, and the relationship between the elements was displayed by knowledge mapping. **[Result/conclusion]** The experimental results show that the F1 values of news elements except Drug are above 80%, which indicates that the transfer learning model can achieve fine recognition effect. Moreover, the knowledge map of associated data can intuitively display the key elements and main contents of news. In conclusion, the method proposed in this paper can effectively identify elements in COVID-19 news, thus it can help readers obtain important information from the news accurately and efficiently.

Keywords: multi-task　transfer learning　COVID-19　news elements identification　named entity recognition　cold start